CMPT 353 D1: Computational Data Science

Term Project

13/08/2021

We Love Cycling

Daniel Wang	301413271
Fitzpatrick Laddaran	301282987
Hong Cung Quang	301417603

Instructor:	Greg Baker
Teaching Assistant:	Ali Arab
Teaching Assistant:	Ghazal Saheb Jam

Table of Contents

Project Overview

Inspired by an avid cyclist within our group, the idea behind this project was to determine whether the variation in a cyclist's power exertion affects one's performance throughout a cycling ride. A cyclist's power exertion is measured by the amount of power exerted through the bike's pedals. This data is recorded via a strain gauge located on the cyclists' bicycle. Typically, this is located on one or both crank arms, but can also be in the crank spindle, crank spider, or the pedal itself. Power variation is calculated with this formula:

$$v_{index} = \frac{P_{normalized}}{P_{average}} \quad (1)$$

where v_{index} is the variability index, $P_{normalized}$ is the normalized power, and $P_{average}$ is the average power. $P_{normalized}$ is calculated as follows:

1. Calculate rolling averages throughout the entire cycling ride.
2. Calculate the fourth power of the rolling averages.
3. Calculate the average of the values in the previous step.
4. Calculate the fourth root of the value in the previous step.

This calculation is based on Aart Goossens'—a data engineer—description of normalized power, which ultimately was adapted from the *Training and Racing with a Power Meter* by Hunter and Allen[1, 2]. Effectively, v_{index} measures how smooth a cyclist's power output is throughout a ride.

Additionally, an elevation scoring index was calculated with the following formula:

$$s_{index} = \frac{e_{gain}}{d_{total}} \quad (2)$$

where s_{index} is the elevation score, e_{gain} is the total elevation gain over the course of the ride, and d_{total} is the total distance of the ride. We calculated s_{index} to find correlations between power variation and different elevation profiles. This is important because we believe that elevation profiles should affect one's overall performance, that is: a hillier ride is likely to be a harder ride and hence one's performance should be weaker in comparison to a flat ride of the same distance.

In the following pages, we first discuss the data gathering and cleaning process. We then provide an overview of the type of analyses we have conducted, and then proceed with the conclusions. Additionally, we list some of the limitations that we have encountered while completing this project.

[1] See https://medium.com/critical-powers/formulas-from-training-and-racing-with-a-power-meter-2a295c661b46.
[2] See https://www.trainingpeaks.com/blog/power-terminology-for-cycling/.

Gathering the Data

Golden Cheetah is an open-source analytics tool that cyclists can use to analyze the data recorded from their rides. Riders can opt in to share anonymized data to the GoldenCheetah OpenData project. Our data was gathered using their API.

Cleaning the Data

The original uncleaned dataset contains many anomalies, which could have originated from various sources. Such sources include:

- sensor malfunction and inconsistency,
- poor cycling routes that contained many stops, and
- unpredictable cycling behaviours.

Therefore, the data was cleaned under these conditions:

- removed rides with incomplete altitude, distance, or power data,
- removed records where idling periods were 150 seconds or greater,
- removed rides with average speeds less than 10 and over 60 kilometers/hour, and
- removed rides that were less than 5 km and over 300 km.

These conditions eliminated situations where:

- significant fluctuations in measured values,
- extreme measured values,
- a rest stop, and
- prolonged downhill riding was observed.

Evidently, the final dataset captured the entire ride while avoiding segments of data that did not help with our research. The cyclist's performance is based on segments of data in which the cyclist is consistently pedalling, or small segments of data in which the cyclist is not pedalling but is consistently moving. Therefore, we keep records where the cyclist may be cruising or on a downhill trajectory.

The final dataset contained 28 unique cyclists, totalling to 4116 rides. This averages out to 147 rides per cyclist.

Analysis and Findings

Given the large dataset, conducting our analyses required us to partition the dataset into an appropriate format. Using Spark, the 28 cyclists were put into small groups of three or four. Additionally, a unique identifier was provided for every record in every cyclist's ride. This allowed us to keep track of which record belonged to which cyclist. For every group of cyclists, all the data for the cyclists in that group were then put into one data frame. Finally, the calculation for v_{index} and s_{index} were done on this data frame. Other statistics of interest were also calculated. These statistics include the total time and distance of the ride, the average power exerted by the cyclist throughout the ride, and the average altitude of the course.

To illustrate, here is an image of a data frame as described above:

cyclist_id	file_name	total_time_sec	total_dist_km	avg_power	avg_alt	v_index	s_index
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_01_14_48_14revised.csv	12752	84.506	196.518	243.593	1.267	12.1719
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_04_13_45_56.csv	2193	7.97832	245.742	734.571	1.207	66.43
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_06_15_31_55.csv	6998	56.7788	205.392	130.933	1.217	10.1552
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_07_15_03_11.csv	9117	67.6581	214.753	132.519	1.252	12.7228
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_10_12_34_00.csv	4498	38.671	234.428	1272.87	1.219	18.991
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_11_09_32_10.csv	10903	70.017	179.493	1471.44	1.289	23.1087
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_15_15_05_14.csv	10651	82.8242	180.64	104.015	1.168	9.03601
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_18_08_38_56.csv	23456	113.199	166.578	1323.26	1.432	23.2158
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_19_09_53_59.csv	18206	129.887	196.626	1256.02	1.2	19.3861
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_20_15_18_10.csv	10172	87.4868	203.155	194.601	1.126	8.53843
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_21_15_22_56.csv	7157	57.8449	183.011	85.669	1.166	7.36798
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_22_15_03_45.csv	10909	85.5793	168.477	63.863	1.206	6.32396
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_23_08_02_37.csv	11259	84.1755	158.684	106.318	1.381	9.75581
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_02_27_15_09_01.csv	9684	83.4082	220.75	102.381	1.165	9.30364
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_03_01_14_42_53.csv	12349	103.322	211.853	177.368	1.234	12.1833
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_03_02_14_59_41.csv	11623	88.854	182.154	95.176	1.178	8.79645
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_03_04_10_13_00.csv	7550	52.9397	238.006	104.93	1.26	16.3507
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_05_03_15_41_33.csv	8495	71.4673	186.8	117.212	1.186	7.47475
000c6417-e1e4-497e-89e6-bb21e17ec355	2017_05_04_08_17_38.csv	16572	113.851	177.903	251.833	1.273	12.1949
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_08_30_05_53_53revised.csv	3391	9.54055	240.218	395.6	1.076	10.6074
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_09_03_14_10_04revised.csv	10211	49.2405	234.531	388.829	1.007	2.55887
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_09_06_05_59_55.csv	4023	12.2464	183.715	379.554	1.039	9.04756
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_09_10_06_45_12.csv	8637	25.1034	248.314	483.709	1.013	20.2443
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_09_12_05_52_40.csv	2720	7.77977	246.137	435.443	1.014	26.9417
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_09_15_05_02_20.csv	6761	21.1967	261.078	359.323	1.006	15.5118
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_09_17_18_41_49.csv	2291	6.43363	255.999	398.258	1.01	20.455
00cab3f8-83c3-4bb8-8c17-aff839b4bed9	2017_09_18_07_32_45.csv	3686	10.7242	250.7	441.241	1.014	23.5915

Figure 1: Data frame after v_{index} and s_{index} calculations.

As seen on Figure 1, every data frame will consist of multiple cyclists. This depends on the number of cyclists partitioned into a specific group. To combat any data misinterpretation, a combination of the cyclist's alias, *cyclist_id*, and one of their rides, *file_name*, has been used as a unique identifier. This means that for every cyclist's ride, we now have some statistics calculated.

At this point, our dataset contained some outliers. This issue originateed from our initial cleaning. The conditions that we have used did not eliminate all the potential anomalies in our data; hence, we eliminated records in which the v_{index} was above two, and where the average altitude was greater than 2500 meters. We set the first condition because the value of v_{index} is generally between one and two, where two is on the extreme side for evaluating variance. The second condition was applied because very few rides averaged greater than 2500 meters.

Beginning our analysis, we created a pair plot to determine if any relationships exist between the variables. These are the results:

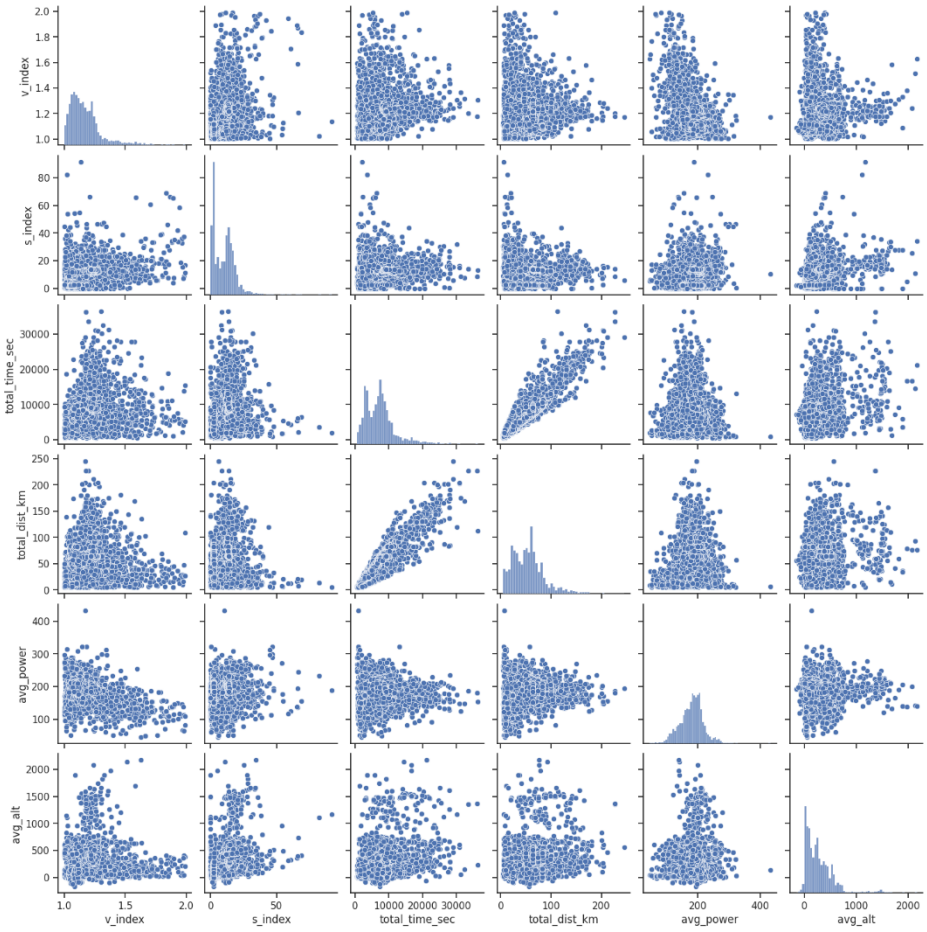

Figure 2: Pair plot for all variable pairings.

Figure 2 does not seem to give a good visualization on the relationships between the variables in our data. The huge number of records seem to obstruct the actual relationships if any exist; however, we can still obtain some general information from the plots. As such, there seems to be a linear correlation between the total time it takes to complete a course, *total_time_sec*, and the total length of the course, *total_dist_km*. Intuitively, this should be true. Looking at the s_{index} column, it seems that there is a vertical trend between all other variables. This indicates that s_{index} may not be a factor that affects other variables.

To solidify our general findings, we also performed a parametric and post-hoc test, namely: the ANOVA and Tukey's Honest Significant Difference (HSD) test. We conducted a parametric test because we have a sufficiently large number of recorded rides, $n > 4000$; however, the lack of normality in the distributions seen in Figure 2 indicates that ANOVA may not be the best statistical tool. Therefore, we considered a non-parametric test, the Kruskal-Wallis test and its corresponding post-hoc test, the Dunn test. The following are the results:

```
Anova testing:
F_onewayResult(statistic=10054.60279892345, pvalue=0.0)
Post_hoc Pairwise_tukeyhsd:
          Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================================
    group1         group2      meandiff  p-adj    lower      upper    reject
-----------------------------------------------------------------------------
       avg_alt      avg_power   -92.5643 0.2441  -213.2492    28.1206  False
       avg_alt        s_index  -262.9569  0.001  -383.6418   -142.272   True
       avg_alt  total_dist_km   -218.252  0.001  -338.9369   -97.5671   True
       avg_alt total_time_sec  7162.0645  0.001  7041.3796  7282.7494   True
       avg_alt        v_index  -271.9428  0.001  -392.6277   -151.258   True
     avg_power        s_index  -170.3926  0.001  -291.0775   -49.7077   True
     avg_power  total_dist_km  -125.6877 0.0356  -246.3726    -5.0028   True
     avg_power total_time_sec  7254.6288  0.001  7133.9439  7375.3137   True
     avg_power        v_index  -179.3786  0.001  -300.0635   -58.6937   True
       s_index  total_dist_km    44.7049 0.8992     -75.98   165.3898  False
       s_index total_time_sec  7425.0214  0.001  7304.3365  7545.7063   True
       s_index        v_index    -8.9859    0.9  -129.6708    111.699  False
 total_dist_km total_time_sec  7380.3165  0.001  7259.6316  7501.0014   True
 total_dist_km        v_index   -53.6909 0.7767  -174.3758     66.994  False
total_time_sec        v_index -7434.0073  0.001 -7554.6922 -7313.3224   True
-----------------------------------------------------------------------------
Kruskal testing:
KruskalResult(statistic=17732.927197006622, pvalue=0.0)
Post_hoc Dunn analysis:
                    avg_alt      avg_power        s_index  total_dist_km  total_time_sec        v_index
avg_alt        1.000000e+00   1.281112e-03   0.000000e+00  2.153390e-186             0.0   0.000000e+00
avg_power      1.281112e-03   1.000000e+00   0.000000e+00  1.224393e-239             0.0   0.000000e+00
s_index        0.000000e+00   0.000000e+00   1.000000e+00  1.243060e-197             0.0  7.552594e-117
total_dist_km  2.153390e-186  1.224393e-239  1.243060e-197   1.000000e+00             0.0   0.000000e+00
total_time_sec  0.000000e+00   0.000000e+00   0.000000e+00   0.000000e+00             1.0   0.000000e+00
v_index        0.000000e+00   0.000000e+00  7.552594e-117   0.000000e+00             0.0   1.000000e+00
```

Figure 3: Results from parametric and non-parametric tests.

From Figure 3, the ANOVA test produced some significant results. The first main interest is the p-value obtained from ANOVA, which is $< \alpha = 0.05$. With Tukey's HSD test, the interesting results indicated that the pairing for s_{index} and v_{index} is inconclusive. Same goes for the pairing of s_{index} with *total_dist_km*, and v_{index} with *total_dist_km.*

With the Kruskal-Wallis test, we also have p-value $< \alpha = 0.05$, so we conducted the Dunn test to determine which pairings were significant. The test tells us that every pairing is significant.

Finally, to answer our original question, we created a density scatter plot of v_{index} against performance, measured as the average velocity.

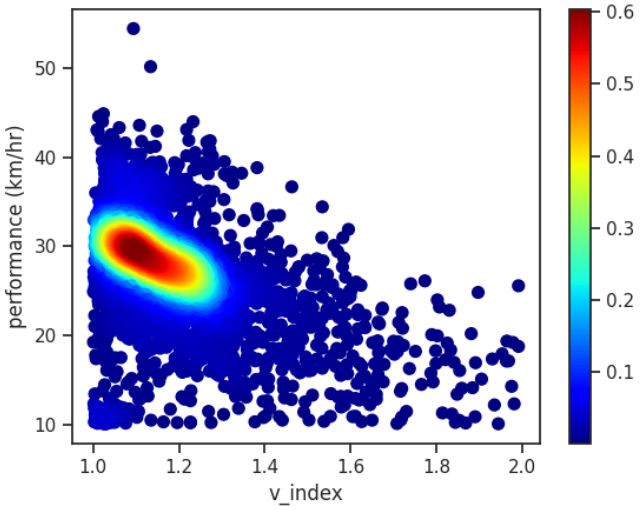

Figure 4: Density scatter plot of v_{index} against performance.

From Figure 4, the density scatter plot shows us that about 60 percent of our rides are on the red-colored surface. This indicates that cyclists who have a v_{index} of about 1.0 to 1.2 are on the scale for average performance. As we see, for values of v_{index} that are closer to 2, cyclists also have poorer performance.

Conclusions

Figure 4 partially answers our original prompt. Cyclists with a lower v_{index} seem to perform better in comparison to those with a higher v_{index}; however, it is slightly difficult to determine whether there really does exist a trend due to large number of observations. Due to the difficulties in visualizing s_{index} as part of the density scatter plot, our group was not able to conclude whether s_{index} has a correlation with v_{index} and performance.

We have created Figure 2 to formalize some trends. This way, we could identify patterns that are not intuitive. As discussed, we see very little patterns.

Finally, Figure 3 aided us by providing potential future improvements to this project. We can conclude which pairings of variables are significant; hence, creating models for a variable should somehow depend on these other significant variables.

Conclusively, our group was satisfied with the project that we have completed.

Limitations and Improvements

In general, the limiting factor that hindered our progress was time. Given everyone's busy schedules, it was hard to coordinate meetings and to motivate progress. If we had more time, we would have done the following:

- research and improve the formulation for s_{index},
- research and deduce other metrics for evaluating performance,
- create a machine learning model that could predict performance based on significant variables,
- improve Figure 2 by creating subsets of the records,
- improve Figure 3 by adding more variables for comparison such as performance,
- improve Figure 4 such that it incorporates the s_{index} as the colored density,
- increase and pick selective sample sizes to ensure representability, and
- refine our data cleaning methods such that it would mitigate anomalies and improve data quality.

In retrospect however, our group should have been more objective-oriented. Instead of leaving things to the last-minute, we should have started earlier and enforced deadlines for various milestones.

Accomplishment Statement

Daniel's Statement

Acquired and cleaned dataset from GoldenCheetah OpenData API. Researched and implemented methods in calculating *elevation variability*.

Fitz's Statement

Researched and implemented methods in calculating *power variability* in cycling performance using standard Python libraries such as pandas. Researched and advised on methods for calculating *elevation variability* in different geographical terrains. Conducted non-parametric statistical tests and post-hoc analysis to deliver various conclusions.

Hong's Statement

Developed the data pipeline and cleaned the data. Created and debugged strava_write.py, strava_analy.py, strava_var_index.py. Implemented statistical analysis.